CLaC at SemEval-2023 Task 3: Language Potluck RoBERTa Detects Online Persuasion Techniques in a Multilingual Setup

Nelson Filipe Costa and **Bryce Hamilton** and **Leila Kosseim**
Computational Linguistics at Concordia (CLaC) Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montréal, Québec, Canada
{nelsonfilipe.costa, bryce.hamilton}@mail.concordia.ca,
leila.kosseim@concordia.ca

Abstract

This paper presents our approach to the SemEval-2023 Task 3 to detect online persuasion techniques in a multilingual setup. Our classification system is based on the RoBERTa-base model trained predominantly on English to label the persuasion techniques across 9 different languages. Our system was able to significantly surpass the baseline performance in 3 of the 9 languages: English, Georgian and Greek. However, our wrong assumption that a single classification system trained predominantly on English could generalize well to other languages, negatively impacted our scores on the other 6 languages. In this paper, we provide a description of the reasoning behind the development of our final model and what conclusions may be drawn from its performance for future work.

1 Introduction

The SemEval-2023 shared task 3 (Piskorski et al., 2023) focuses on 3 subtasks: detecting the news genre (subtask 1), the framing (subtask 2) and the persuasion techniques (subtask 3) used in online news across different languages. Training data provided by the organizers was collected from news and web articles between the years 2020 and 2022 in six different languages, namely English, French, German, Italian, Polish, and Russian. During the testing phase of the shared task, three additional surprise languages were included only as testing data, Georgian, Greek and Spanish, to explore the language-agnostic robustness of the trained models.

In this paper, we focus on the detection of persuasion techniques used to influence the reader of these online news as part of subtask 3 (Piskorski et al., 2023). We did not participate in the news genre categorisation (subtask 1) and framing detection (subtask 2). In total, 23 different persuasion techniques were annotated at the paragraph-level for each article across all the training languages. Such techniques include *loaded language*, *guilt by association* and the *straw man fallacy*. A more detailed description of each persuasion technique is available in the annotation guidelines of the shared task (Piskorski et al., 2023).

We applied a RoBERTa (Liu et al., 2019) based model trained predominantly on the English training dataset enriched with specific instances from the other training languages to the detection of persuasion techniques across all languages. Section 3 provides an overview of our final classification pipeline. In Section 4, we provide a detailed description of the experiments and data augmentation techniques that led to our final model decision. However, our assumption that a predominately English trained model would be able to generalize well and even outperform models trained on languages specific to the prediction task was proven wrong by the official leaderboard results of the shared task as further discussed in Section 5.

With the exception of English and two of the three surprise test languages, Greek and Georgian, where we were able to attain significantly higher scores than the baseline, our final model either performed worse or only slightly better than the baseline. All of the code used in the implementation of the models described in this paper is made available on GitHub[1].

2 Background

The massively widespread of misinformation to deliberately influence opinions and beliefs through the use of rhetorical and psychological persuasion techniques, commonly referred to as *fake news*, has become a major societal problem capable of even swinging elections (Zellers et al., 2019). It has, thus, become of paramount importance to develop systems capable of automatically detecting

[1] https://github.com/CLaC-Lab/Persuasion-Techniques-Detection

such persuasion techniques (Da San Martino et al., 2019).

With the advent of the Transformer (Vaswani et al., 2017) architecture and ensuing pre-trained large language models, such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), the automatic detection of misinformation in online news has become feasible with high performance and hence an active area of research (Zellers et al., 2019). The detection of propaganda techniques in news articles was in fact the goal of the SemEval-2020 shared task 11 (Da San Martino et al., 2020).

Based on the approach followed by the best participants at the SemEval-2020 shared task 11 (Jurkiewicz et al., 2020) and the recent successes of Transformer based pre-trained large language models, we based our system for the SemEval-2023 shared task 3 (Piskorski et al., 2023) on RoBERTa (Liu et al., 2019).

3 System Overview

The goal of the task is to identify all persuasion techniques (if any) employed at the paragraph-level in online news articles. Therefore, our model takes as input an entire paragraph and classifies it as exhibiting none, one or multiple persuasion techniques as output. Figure 1 shows an overview of the classification pipeline we employed for this subtask. As Figure 1 shows, our classification system is composed of a RoBERTa-base[2] model with a feedforward layer of 23 nodes, one for each of the annotated techniques. We used a sigmoid activation function at the output layer and we considered the entire task a 23-class multi-label binary classification problem.

Since the RoBERTa-base model was pre-trained on the English language exclusively, for all the other training languages we first had to translate the datasets into English before feeding them to our model. We did this using the Googletrans[3] API. We then augmented the datasets of each language (see Section 4.2). Note that for each language we could also have used a different RoBERTa based model pre-trained on that specific language, but instead we opted to first translate the different languages into English and then use RoBERTa-base. This decision was motivated by our experiments with the development set (see Section 4.3).

[2] https://huggingface.co/roberta-base

[3] https://py-googletrans.readthedocs.io/en/latest

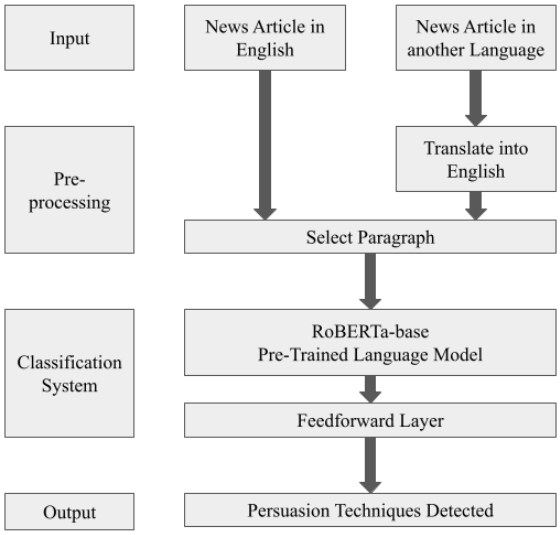


Figure 1: Schematic overview of our final classification pipeline for the detection of persuasion techniques in online news articles.

4 Experimental Setup

Before opting for our final classification system configuration, we experimented with different augmentation techniques, we considered different language models and we explored training our models with specific language datasets to classify the persuasion techniques across all languages. In this section we will describe in detail our experiments and the decision process that led to our final model.

4.1 Data Splits

Data for each language was made gradually available during three different phases of the shared task. In the first phase, participants had only access to the annotated training datasets of English, French, German, Italian, Polish, and Russian; in a second phase, participants had additional access to the annotated development datasets of the same languages; and in the third and final phase, the annotated test datasets with hidden labels of the same languages and of the surprise Georgian, Greek and Spanish were made available.

During the first phase, we randomly divided the training datasets into three different splits: 70% for training, 10% for validation and 20% for testing. During the second phase, we randomly split 80% of the original training datasets for training and 20% for validation, while using the newly available development datasets for testing. Note that the development set of each language was around 25-35%

the size of its corresponding training dataset. This meant that for each language we still had around 70% of the combined data for training, 10% for validation and 20% for testing. Finally, during the third and final phase, we merged both training and development datasets into single datasets for each language, which we then randomly split 80% for training and 20% for validation, and used the new test datasets with hidden labels to submit our final classification predictions.

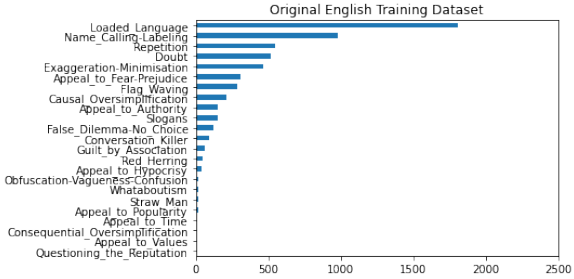
Throughout all phases, we used the same set-up: we used a batch size of 8 and kept the training to a maximum of 10 epochs with an early stop call after 2 epochs without improvement on the validation set. Since we applied a sigmoid activation function at the output layer of the multi-label binary classification system, we used a binary cross entropy loss function across all experiments.

4.2 Data Augmentation

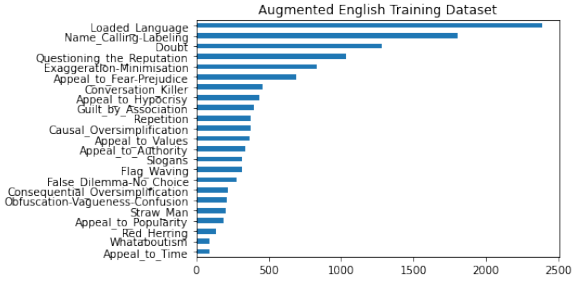
As described above, each language had its own training, development and test datasets. However, the distribution of persuasion techniques was not the same across languages. In fact, certain techniques were even missing in some languages. Figure 2a shows the distribution of the original English training dataset. As shown in the figure, the dataset includes no instances of the *Appeal to Time*, *Consequential Oversimplification*, *Appeal to Values* and *Questioning the Reputation* techniques. Conversely, there is an unbalanced over-representation of certain techniques, such as *Loaded Language* and *Name Calling Labelling*. This was a recurring problem across all datasets.

Therefore, in order to have a representative number of samples of all persuasion techniques, we first enriched each dataset with instances of missing or under-represented persuasion techniques by translating the instances from other languages into the original dataset language. This approach was motivated by the already successful use of automatically translated data in subjective-oriented natural language processing tasks (Banea et al., 2008).

We considered instances to be under-represented if they appeared less than 5% of the total number of labelled instances. However, due to the multi-label nature of the task, some of the newly added instances were also jointly labelled with already over-represented techniques in the original datasets. Thus, after augmenting the original datasets, we randomly dropped instances of the



(a) Original English Training Dataset.



(b) Augmented English Training Dataset.

Figure 2: Distribution of persuasion technique instances between the original and the augmented English training dataset.

over-represented techniques that did not co-occur with the under-represented ones. We considered instances to be over-represented if they appeared more than 20% of the total number of labelled instances in the augmented dataset.

Finally, we dropped all instances without a label from all the datasets used for training our models. These were the mostly over-represented instances across all languages and the shared task did not evaluate our ability to predict them. Figure 2b shows the final distribution obtained after augmenting the original English training dataset.

4.3 Model Selection

During the model selection stage, we considered different models in the BERT (Devlin et al., 2019) family. We considered only these models because of their ability to learn a bidirectional representation of a sentence given their pre-training objective. This makes them particularly well-suited for the downstream task of detecting persuasion techniques at the paragraph-level. This is different from traditional LSTM recurrent neural networks (Hochreiter and Schmidhuber, 1997) that see the words one after the other and from autoregressive models such as in the GPT (Radford et al., 2018) family.

Language \ Model	RoBERTa	CamemBERT	German-BERT	Multilingual-BERT
English	**0.34942 ± 0.00358**	-	-	0.29700 ± 0.00568
French	**0.41356 ± 0.00467**	0.17875 ± 0.02104	-	0.38295 ± 0.01085
German	**0.34643 ± 0.02028**	-	0.22829 ± 0.02667	0.31291 ± 0.01997
Italian	**0.46358 ± 0.00502**	-	-	0.43208 ± 0.01718
Polish	**0.29224 ± 0.00909**	-	-	0.24805 ± 0.01031
Russian	**0.32399 ± 0.00671**	-	-	0.31832 ± 0.01527

Table 1: Comparison of the F1 micro-averaged scores obtained by the different models when trained and tested on the same language. These are average scores obtained after training and validating each model 5 times on different partitions of its training set and testing it on the corresponding development set. Results in bold represent which model obtained the best score in each given language. For the RoBERTa based model, all languages with the exception of English were first translated into English.

As shown in Table 1, we experimented with four different models. We considered RoBERTa-base on all languages after translating them to English (see Figure 1), as well as the French based CamemBERT-base[4] (Martin et al., 2020) on the French language only, the German based German-BERT-base[5] on the German language only and, finally, the Multilingual-BERT-base[6] natively on all languages. All four models were trained and tested on the same language datasets after being enriched with instances of their missing or under-represented techniques from the other languages (see Section 4.2). They were trained exclusively on data from the training datasets and tested on the separately released development datasets for their corresponding training language (see phase 2 in Section 4.1).

Based on the results shown in Table 1, the RoBERTa based model outperformed all other models. Surprisingly, it even outperformed the CamemBERT based model on French texts and the German-BERT on German texts. Based on these results, we further experimented with other variations of RoBERTa; namely, RoBERTa-large[7] and XLM-RoBERTa-base[8] (Conneau et al., 2020). However, these larger models took significantly more time to train and did not provide substantial improvements in performance, so they were ruled out from further experiments and we chose RoBERTa-base as our final model (see Figure 1).

[4] https://huggingface.co/camembert-base

[5] https://huggingface.co/bert-base-german-cased

[6] https://huggingface.co/bert-base-multilingual-cased

[7] https://huggingface.co/roberta-large

[8] https://huggingface.co/xlm-roberta-base

4.4 Training Language Selection

After selecting our final model, we additionally explored if we could train a single model on a specific language and then generalize it to predict persuasion techniques on all other languages after translating their testing data into the training language. These results are shown in Table 2. Row-wise, the table shows the performance of a model trained on a given language training dataset across the different development datasets of the different languages. Column-wise, the table shows how well the models trained on different languages were able to perform on a given language development dataset.

As Table 2 shows, our RoBERTa-base model performed particularly well across all languages when trained on English and it performed especially well on both Latin languages, French and Italian. Based on these results, we assumed that there could be an etymological explanation behind the structure of the English language that led to these results.

5 Results

Based on the results of Table 1 and Table 2, for our official submission, we trained our single RoBERTa-base classification system described in Section 3 on the merged English training and development datasets following the augmentation techniques discussed in Section 4.2. We then translated all of the provided test datasets into English before generating our final classification predictions on each language. The official performance results of our system are shown in Table 3, along with the baseline score and the score obtained by the best performing system on each language.

As Table 3 shows, the preliminary results obtained with the training and development datasets

Training \ Testing	English	French	German	Italian	Polish	Russian
English	**0.34942 ± 0.00358**	**0.42384 ± 0.01961**	**0.35792 ± 0.01985**	**0.46926 ± 0.02086**	0.30393 ± 0.01002	**0.34546 ± 0.02682**
French	0.28395 ± 0.02024	0.41356 ± 0.00467	0.35031 ± 0.01105	0.43084 ± 0.01466	0.30533 ± 0.00629	0.30495 ± 0.00667
German	0.27434 ± 0.00461	0.40070 ± 0.00861	0.34643 ± 0.02028	0.42529 ± 0.01232	0.28828 ± 0.01473	0.29738 ± 0.01444
Italian	0.32418 ± 0.01468	0.41212 ± 0.00984	0.34655 ± 0.00472	0.46358 ± 0.00502	**0.31256 ± 0.01007**	0.33540 ± 0.01400
Polish	0.29407 ± 0.01155	0.39555 ± 0.01571	0.32058 ± 0.00601	0.42647 ± 0.01038	0.29224 ± 0.00909	0.30859 ± 0.01789
Russian	0.29617 ± 0.01187	0.40199 ± 0.01562	0.32945 ± 0.00749	0.44487 ± 0.00942	0.28886 ± 0.01024	0.32399 ± 0.00671

Table 2: Comparison of the F1 micro-averaged scores obtained by our RoBERTa-base model when trained on a language and tested on all other languages. These are average scores obtained after training and validating each model 5 times on different partitions of its training set and testing it on the corresponding development set. Results in bold represent which training language obtained the best score in each test language. All languages with the exception of English were first translated into English.

Language	Baseline	Our Score	Best Score
English	0.19517	0.30933	0.39029
French	0.24014	0.23903	0.46869
German	0.31667	0.24795	0.52006
Italian	0.39719	0.31310	0.56480
Polish	0.17928	0.19036	0.43037
Russian	0.20722	0.19285	0.38682
Georgian	0.13793	0.27053	0.45714
Greek	0.08831	0.15630	0.27827
Spanish	0.24843	0.26667	0.38106

Table 3: Comparison of the final F1 micro-averaged scores obtained by our classification system, the best corresponding classification system in the shared task and the baseline in each given language.

(see Table 2) which led us to develop an English-only trained model, were not representative of the official performance obtained with the final test datasets of the shared task. Considering only our results on the languages for which a training set was provided, the model was able to significantly outperform the baseline only on English. In fact, it performed worse than the baseline in French, German, Italian and Russian. This could be a result of different label distributions between the training datasets and the final test datasets. However, the assumption that a predominantly English trained model could generalize well to the task of detecting persuasion techniques across different languages than training a model natively on the same language of the predictive task, turned out to be wrong.

On the three surprise languages, our classification system performs substantially better than the baseline on Georgian and Greek, while only slightly better on Spanish. Given that we did not have access to any training dataset in the Georgian and in the Greek family of languages, these results may indicate that although our assumption was wrong, it may not have been entirely unfounded. Thus, a model trained in a Latin language, such as French or Italian, could have perhaps led to better results in the Spanish language.

6 Conclusion

In this paper we described our approach to the SemEval-2023 Task 3 to detect online persuasion techniques in a multilingual setup. We applied a single classification system based on the RoBERTa-base model trained predominantly on the English language to detect persuasion techniques across 9 different languages. Based on the results obtained with the development set, we assumed that our English trained model could generalize well to every other language and, hence, we did not develop language specific models. Given the actual test set, the model was only able to significantly outperform the baseline for English and the two surprise, Georgian and Greek, for which no previous training data was available.

In the future, we plan to explore the use of other language models, such as ALBERT (Lan et al., 2019), and to evaluate how our final classification system would have performed if trained on the same language of the prediction task. We also plan on exploring a cascading model of classification systems for each persuasion technique on a one-vs-all scenario. Finally, given the different structure of the different languages, we would also like to investigate which techniques are more frequently classified correctly and which techniques are more frequently classified incorrectly on each language. This may lead to practical insights on how subjective information may be lost using automatic machine translation.

Acknowledgements

The authors would like to thank the anonymous reviewers for their feedback on the previous version of this paper. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual Subjectivity Analysis Using Machine Translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127–135, Honolulu, Hawaii. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NIPS)*, Vancouver, British Columbia, Canada. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations . *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving Language Understanding by Generative Pre-Training.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, California, USA. Curran Associates, Inc.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NIPS)*, Vancouver, British Columbia, Canada. Curran Associates Inc.